Breakout Group 2 (QA/QC) Workshop summary points

Participants: Baker, Bliss, del Grosso, Denning, Dilling, Goebel, Gu, Heath, Jarvis, Kozyr, Krause, Masarie
CCIWG: Murdoch

Linda Heath, *Rapporteur*

Strategy – near vs long term needs

- There are near-term data management needs of PIs who are currently collecting data. Use existing tools/data centers/PIs to manage this data. Efforts must begin soon.
- Meanwhile, a vision of a data mgmt system to be implemented in the longterm (3 yrs) will be documented. Longterm archival storage of new and of older important data and data products must be addressed.

Data mgmt system design (1 of 2)

- Flexible system that includes using existing data centers, PIs for data, and a centralized data center.
- Concentrate on near term needs for now (current PIs) using existing tools/centers.
- Document system requirements needed for long term. Review document before building system.
- Need complete and clear metadata.

Data mgmt system design – (2 of 2)

- Accommodate all data types: Raster up to 4D, Vector,Point,Time series, Survey (social science)

QA/QC, Data Quality Act?

- QA/QC is needed at all levels, using standard methods.
- Quality/quality control occurs at the producer level. We recommend no “office” of QA/QC at NACP.
- Some accuracy issues are use-specific. (Ex: map accuracy for highly defined points vs vague locations). QA/QC in metadata may need to list different uncertainties for different science questions.
- Versioning of data. Need to preserve old versions while including improved datasets (perhaps due to gapfilling). Agencies may not do this. Ex: National Elevation Dataset-DEMS.

Identify nearterm and longterm activities (pilot/prototype)—these are nearterm

- Need to know what datasets are needed and when.
- Questionaire for NACP investigators as to what data they are going to need and for what purpose. Goal is to help scientists find what they need AND to help define metadata needs in terms of QA/QC (see point 3 previous slide).

Identify nearterm activities (con't)

- NACP web site explaining all NACP investigations going on now to help people find each other and make human contacts.
- Hire a NACP data worker bee-queen hybrid. First task: Person to review existing metadata tools for applicability, facilitate cataloging of data And they need to start soon. This paid position interfaces with DMC and is responsible for day to day work.

Existing tools

- **Mercury** is a good cataloging tool. Main limitation is that you can't search inside the file. So you have to download the entire file. Use with web based metadata form: OME.
- Look at GCIP/Gewex program as example as how to handle a variety of scales and capabilities. Used by Europeans.
- Look at OpenDAP. Allows subsetting of files and visualization. Threadss is the corresponding Open Source cataloging tool.

Existing data centers that could be utilized at least in the near term for data management?

- Do we already have data that could function equivalent to the ones in CarboEurope-IP? Examples: EROS, Goddard-DAAC, ORNL-CDIAC,CIESEN, USDA
- We can also have PIs for data, existing data centers, and centralized data center.

Creating needed data layers

- If centralized GIS services are needed: some existing data centers (eg Eros or ORNL) can do some of reformatting/GIS work that is needed cost effectively and with high quality QA/QC. Some of it might even be funded through current channels, people just need to be told to do the work.

Identify longterm activities

- Flexible system for either PI to handle data service or to submit for centralize.
- Work with international partners.
- Want the system to be able to provide subsets of data.

How to exert oversight and mgmt of the NACP DMP

- Need to establish a standing data management committee to:
  - oversee and facilitate data management
  - interface with big data providers/centers,
  - interface with PIs,
  - and help determine and decide users and user needs.
  - This committee will report..or be a subset of? to NACP SSG
- One position in NACP, as a data queen-worker bee type.

Breakout questions, and summary of answers

- 1. Can we identify what measures of uncertainty and bias should be reported with data and data products?
- 2. Can we adopt existing guidelines for evaluating and expressing uncertainty of data (eg ANSI/NCSL)?
- 3. Can we establish a protocol for including QA/QC data with measurement data submissions?

Questions (cont) and responses

- 4.Can we produce QA/QC information in a user-friendly (useable) format?
- 5. Can we establish a mechanism for documenting/summarizing the QA/QC status of all datasets?
- 6. Can we establish a mechanism for documenting summarizing known data problems?

The response to all these questions is yes! And we should do all of them for good QA/QC.

What hasn’t been addressed?

- Measures of success.
- Long term archive, and permanence.